DyaPlex: Full-Duplex Speech-Motion Model for Dyadic Interaction

Koki Nagano[1†], **Hongyu Liu**[1,2*†], **Seonwook Park**[1], **Tianye Li**[1]
Amrita Mazumdar[1], **Christian Jacobsen**[1], **Shengze Wang**[1], **Michael Stengel**[1]
Rajarshi Roy[1], **Ka Chun Cheung**[1], **Simon See**[1], **Shalini De Mello**[1]
[1]NVIDIA, [2]HKUST

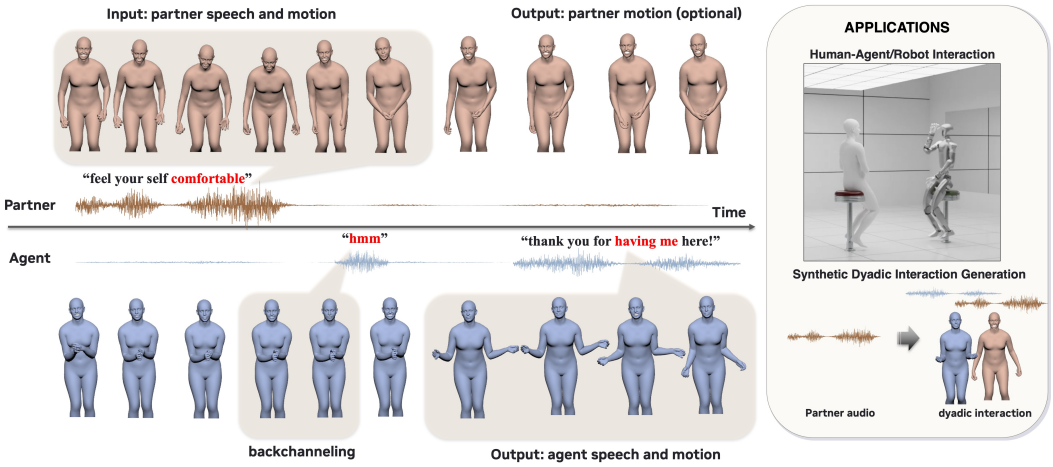

Figure 1: **DyaPlex** is a causal, full-duplex speech-motion model that simultaneously speaks and listens to a partner while perceiving a partner motion and generating agent's motion. Our model could be applied to applications, such as, dyadic interactions with a human user and agent/robot (right) as well as generating synthetic speech-motion dyadic interaction data.

Abstract

We present DyaPlex, a streaming, full-duplex speech-and-motion model designed for dyadic interaction. To capture the continuous and reciprocal nature of human communication, this full-duplex capability empowers the agent to simultaneously perceive and generate both speech and physical motion in a streaming fashion. At its core, our method leverages the strong priors of a foundational full-duplex speech model and integrates a novel motion pathway, thereby achieving fully synchronized multi-modal interaction. Specifically, we design a dual-tower Transformer architecture that preserves the zero-shot conversational reasoning of a frozen base speech model while constructing a deeply coupled, streaming motion pathway. By introducing a unified dyadic token interleaving mechanism and guiding cross-attention via a time-aligned speech-motion RoPE, our model effectively aligns autoregressive motions with rich latent speech features. Trained on the 4,000-hour Seamless Interaction dataset, our model effectively captures cross-speaker dependencies and establishes new state-of-the-art performance across both monadic and dyadic human interaction benchmarks.

*Part of the work was done during an internship at NVIDIA.
†Joint first authors.

Preprint.

1 Introduction

In natural dyadic interaction, listening and expressing are never discrete alternating turns, but rather a highly synchronized, streaming process encompassing both speech and physical movements. Although recent *full-duplex speech* models [4, 30] have facilitated fluid verbal exchanges, genuine interaction is fundamentally multi-modal. Establishing true social presence demands the tight coupling of speech and motion. This *full-duplex speech-and-motion* mutual perception—where participants simultaneously react to verbal and physical cues—is essential for capturing the intricate dynamics of spontaneous encounters. These capabilities are foundational for next-generation Embodied Conversational Agents (ECAs) to navigate complex social nuances, directly enabling responsive human-robot interaction and high-fidelity synthetic dyadic data generation (e.g., Fig. 1).

Recently, several methods have advanced dyadic interaction by incorporating both speech and motion (see Table 1). However, they still fall short of true *full-duplex speech-and-motion*, which requires simultaneously perceiving the partner's multi-modal cues (speech and motion) while generating the agent's reactive responses in a streaming, causal manner. Specifically, prior works such as Audio2Photoreal (A2P) [24] and DyaDiT [28] rely on non-causal diffusion models; while achieving high visual quality, they are restricted to offline generation and are fundamentally precluded from streaming interaction. Similarly, ViBES [39] employs an LLM backbone to jointly model speech and motion, but utilizes a non-causal motion tokenizer and remains entirely blind to the partner's physical movements. Conversely, concurrent causal methods like SARAH [25] and MIBURI [22] support streaming generation but suffer from severe perception deficits. SARAH perceives only limited spatial signals—such as the partner's 2D floor position—ignoring the semantic richness of actual body gestures. MIBURI successfully achieves full-duplex interaction in the *speech* domain, yet its motion generation remains entirely *monadic* (single-person). Because MIBURI cannot receive the partner's motion as input, it generates agent actions conditioned solely on speech.

Table 1: **Capability comparison across representative systems for human–AI dyadic interaction.** The *Full-duplex* column reports the modality in which the system simultaneously *perceives the user* and *generates the agent* in a streaming, causal manner. **DyaPlex** is the only method whose motion pathway is full-duplex (perceives user motion *and* streams agent motion), in addition to inheriting full-duplex speech from its backbone. Furthermore, it is trained on a corpus roughly 20–500× larger than the corpora used by prior work. † SARAH perceives user's 2D floor position but not gestures.

Method	Perception		Generation		Dyadic	Causal	Full-duplex	Training Data
	Partner Speech	Partner Motion	Agent Speech	Agent Motion				
SARAH [25]	Yes	No†	No	Yes	Yes	Yes	No	Embody 3D (50 h)
A2P [24]	Yes	No	No	Yes	Yes	No	No	Custom (8 h)
DualTalk [29]	Yes	Yes	No	Yes	Yes	No	No	DualTalk (50 h)
ViBES [39]	Yes	No	Yes	Yes	No	No	No	Converse 3D (1,000 h)
MIBURI [22]	Yes	No	Yes	Yes	No	Yes	Speech	BEAT2 (70 h)
DyaDiT [28]	Yes	Yes	No	Yes	Yes	No	No	Seamless (182 h)
DyaPlex (Ours)	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Speech & Motion**	**Seamless (4,000 h)**

As highlighted in Table 1, the absence of a unified, causal perception-generation loop across both modalities fundamentally breaks the natural communication flow. Consequently, systems without partner motion perception struggle to produce reciprocal behaviors jointly driven by speech and motion. To illustrate this critical bottleneck, consider conversational dynamics like *subconscious mirroring* and *silent backchanneling*. In natural encounters, a listener often continuously copies a speaker's posture or provides sustained motion feedback (e.g., vigorous nodding) *while* actively listening, without interrupting the speech flow. Simultaneously, the speaker perceives these motion responses and dynamically adjusts their own behavior, creating a truly natural, closed-loop interaction. Traditional frameworks—forced to act as delayed turn-takers or conditioned solely on speech—remain completely oblivious to these concurrent motion cues. In such scenarios, methods failing to simultaneously perceive both the partner's speech and motion break down entirely.

To bridge this gap, we present **DyaPlex**, the first streaming, full-duplex speech-and-motion model for dyadic interaction. By seamlessly conditioning the agent's actions on concurrent multi-modal cues (speech and motion) at the frame level, it enables deeply coupled social behaviors unsupported

by prior models. Our method employs a dual-tower Transformer architecture that adds motion modeling capabilities to a full-duplex speech backbone (e.g., PersonaPlex [30]) while preserving its conversational reasoning. Specifically, we utilize the frozen speech model as the *speech tower*. This tower processes dyadic audio and extracts rich, per-layer residual-stream hidden features that capture the linguistic and prosodic nuances of the conversation. These features are then injected into a trainable *motion tower* via cross-attention. Crucially, the motion tower embeds both participants' physical motions within a unified, autoregressive sequence. This design enables the motion features of the partner and the agent to interact deeply through self-attention, while the cross-attention mechanism ensures these representations attend to their corresponding speech signals. This dual-attention synergy not only aligns the generated motion with the speech but also fully exploits the rich prior knowledge of the foundational speech model (see Fig. 1). Moreover, to ensure precise temporal alignment between the speech and motion streams, we introduce cross-attention with time-aligned RoPE [32]. By injecting explicit relative temporal distances, this design effectively synchronizes the speech-motion modalities and prevents the cross-modal mapping from degenerating into time-agnostic, fixed speech feature retrieval. Finally, because the motion tower maintains strict causality—consistent with the foundational speech backbone—DyaPlex inherently supports seamless streaming generation.

Trained on the 4,000-hour Seamless Interaction [1] dataset–a scale roughly 20 to 500 times larger than previous dyadic motion corpora–DyaPlex captures the vast diversity and high-order dependencies inherent in natural human coordination. Extensive evaluations demonstrate that our model establishes new state-of-the-art performance across both monadic and dyadic human interaction benchmarks, yielding unprecedented realism, synchrony, and socially coherent behaviors.

The primary contributions of this work are summarized as follows:

- We introduce DyaPlex, the first full-duplex model capable of streaming, simultaneous perception and generation of both speech and motion for human-agent interaction.
- We propose a novel dual-tower architecture that deeply couples a trainable motion pathway with a frozen full-duplex speech model. This is achieved via unified dyadic token interleaving and a time-aligned speech-motion RoPE to ensure precise cross-modal synchronization.
- Extensive evaluations show our model establishes new state-of-the-art results on both monadic and dyadic benchmarks, unlocking novel applications such as responsive social robotics and scalable synthetic data generation.

2 Related Work

2.1 Single-Person and Co-Speech Motion Generation

Early 3D motion synthesis focused on single-person generation from text [19, 33, 9, 38] or speech [8, 15, 36, 41]. While recent models [16, 21, 17] significantly advance fine-grained audio-motion synchronization, they remain strictly monadic. Concurrent attempts to unify multi-modal inputs, such as ViBES [39], fuse speech and motion inside a single backbone. However, ViBES is non-causal for tokenizer and hence non-streaming, and forcing all modalities into a shared backbone via self-attention severely limits the motion context window.

2.2 Human Interaction and Reaction Generation

Beyond single characters, prior works forecast multi-person trajectories [11], synthesize interactions from text [34, 14], or generate physically plausible action-reaction flows (e.g., dodging, handshaking) conditioned on a partner's motion [2, 10, 35, 6]. While excelling at spatial coordination, these methods operate in silent, non-conversational environments. Focusing strictly on pure motion-to-motion kinematics, they struggle to generalize to conversational dyadic settings where non-verbal social nuances and continuous spoken dialogue are deeply intertwined.

2.3 Dyadic Conversational Human Interaction

The most relevant domain to our work is dyadic conversational interaction, requiring the simultaneous coordination of verbal and physical behaviors. Early models focused on localized listener

feedback [23]. Subsequent full-body dyadic models prioritize synthesis quality over real-time interactivity, relying on offline audio processing [24], bidirectional recurrent networks [29], or non-causal diffusion frameworks (e.g., DyaDiT [28]), which fundamentally precludes streaming applications.

Recent concurrent work proposed causal methods to achieve real-time performance, but exhibit critical multi-modal perception gaps. SARAH [25] streams dyadic motion but only perceives the user's 2D floor-projected position, entirely ignoring the semantic richness of upper-body gestures. Conversely, MIBURI [22] achieves full-duplex speech interaction, yet its gesture generation remains entirely monadic; without perceiving the partner's motion, it fails to produce visually driven reciprocal behaviors. DyaPlex addresses these limitations directly. By operating on a dyadically-interleaved motion stream within a novel dual-tower architecture, it is the first to achieve a true full-duplex loop—simultaneously perceiving partner full-body motion and speech and streaming multi-modal responses causally.

3 Preliminaries

DyaPlex couples two components: an off-the-shelf full-duplex speech model (PersonaPlex [30]) that remains frozen, and a body-part-aware streaming RVQ-VAE motion tokenizer that we pre-train independently. We summarize their key features below.

3.1 PersonaPlex: Full-Duplex Speech Tower

We use PersonaPlex [30], a causal Transformer built on Moshi [4], which is a full-duplex speech-language architecture featuring a hidden dimension of $d_s = 4096$ and $L_s = 32$ layers (see Fig. 2 (b)). Both speakers' speech is tokenized at f_s=12.5 Hz with the Mimi neural speech codec [4]. At frame t, PersonaPlex receives a 17-way interleaving of text and dyadic speech codebooks. At the embedding layer, these 17 per-codebook embeddings are summed into a single d_s-dim vector per frame. We freeze PersonaPlex throughout our training. Crucially, for every transformer block ℓ=1, . . . , L_s, PersonaPlex exposes its post-block residual-stream hidden states $\mathcal{H}_\ell \in \mathbb{R}^{T\times d_s}$, where T is the sequence length. These hierarchical hidden states capture the rich conversational context and are directly consumed by our motion tower (Sec. 4).

3.2 Body-part-based RVQ-VAE Motion Tokenizer

We tokenize body and (optionally) face motion with a body-part-aware RVQ-VAE, adapted from GestureLSM [17]. We retrain the body tokenizer end-to-end on the Seamless Interaction dataset [1] and modify it into a causal streaming architecture: the encoder consumes 25 fps motion as input and outputs tokens at the speech-aligned rate f_m=12.5 fps, while the decoder applies 2× temporal upsampling to reconstruct 25 fps SMPL-X output. Our RVQ-VAE consists of four independent decoders specialized to the upper body, hands, lower body, and face (see Fig. 2 (a)). Each frame is encoded by K=22 codes (18 body codes and 4 face codes) drawn from a shared vocabulary of size V_{mot}=4096, which is partitioned into four disjoint 1024-entry bands. At inference, these codes are routed to their respective decoders to reconstruct SMPL-X [26] and FLAME [13] parameters. Both the encoder and decoder are frozen during the subsequent motion-tower training in Fig. 2 (b). For complete details regarding the adaptation of the causal architecture, the end-to-end streaming training procedure, and the decoding of tokens into 3D meshes, please refer to the supplement.

4 Full-duplex speech-motion model

We frame full-duplex dyadic interaction as time-aligned autoregressive speech and motion generation for both interacting partners via dedicated speech and motion towers. A frozen speech tower (Sec. 3.1) models audio; a trainable motion tower cross-attends to the speech tower's per-layer hidden features through learned key/value projections. Our motion tower generates motion for *both* speakers of the dyad in a single interleaved stream, enabling the agent to both simultaneously *perceive* the partner's body motion and *respond* through its own. Fig. 2 summarizes the design.

Notation. We consider a dyad (A, B) where, without loss of generality, A is the partner and B the agent. Their Mimi audio tokens are $\mathbf{s}^A_{1:T}, \mathbf{s}^B_{1:T}$ at f_s=12.5 Hz; their RVQ-VAE motion tokens are

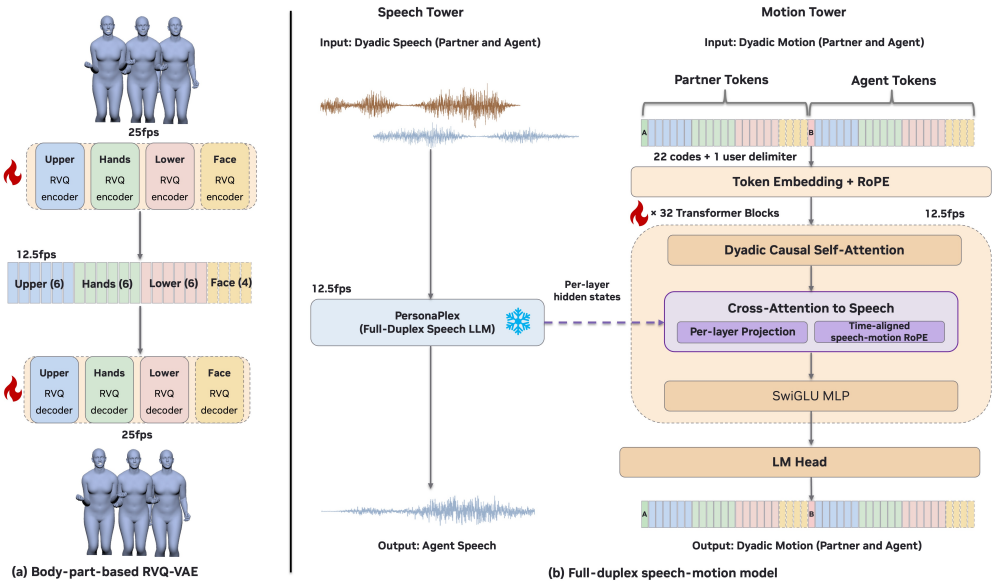

Figure 2: **Architecture overview.** DyaPlex consists of three components: (a) part-aware RVQ-VAE decoders and (b) a frozen speech tower, and a trainable motion tower. The speech tower (PersonaPlex) takes in dyadic speech, emits agent speech autoregressively, and exposes its per-layer residual-stream hidden states $\{\mathcal{H}_\ell\}_{\ell=1}^{32}$. For training, we precompute $\{\mathcal{H}_\ell\}$ once (Sec. 4.1) to serve as cross-attention keys and values for the motion tower. The 32-layer causal motion tower (d_m=1024, h_m=16) operates on a dyadically interleaved stream at 12.5 fps: $[[\texttt{A}], \mathbf{m}_t^A, [\texttt{B}], \mathbf{m}_t^B, \dots]$. Each block applies dyadic causal self-attention on the motion stream, followed by cross-attention to the speech states using a learned projection (h_c=12, d_c=64) and *time-aligned speech-motion RoPE*. The LM head outputs motion tokens (supervised on 18 SMPL-H body codes for our body-only base model) at 12.5 fps, which the RVQ-VAE decoders with 2× temporal upsampling then finally reconstruct into SMPL-X pose parameters at 25 fps.

$\mathbf{m}_{1:T}^A, \mathbf{m}_{1:T}^B$ at f_m=12.5 fps, T is the total number of frames. Each motion frame is a K=22-dim integer vector $\mathbf{m}_t = (c_t^{(1)}, \dots, c_t^{(K)})$ over a shared vocabulary of size V_{mot}=4096.

4.1 Speech hidden-state extraction with a hybrid input prefix

The motion tower conditions on PersonaPlex's per-layer residual-stream hidden states $\{\mathcal{H}_\ell\}_{\ell=1}^{L_s}$ via cross-attention at every block. These hidden states encode the joint conversational context — both speakers' acoustic content, PersonaPlex's inner-monologue text predictions, and the implicit turn-taking dynamics of the dyadic input format. Exposing all L_s=32 layers rather than a single final embedding gives the motion tower's cross-attention access to all intermediate speech representations, not only the final-layer embedding. We pair motion-tower blocks one-to-one with PersonaPlex layers (Sec. 4.2), letting cross-attention learn its own hierarchical speech representations. For training, the causal architecture of PersonaPlex allows us to precompute all hidden states in a single teacher-forced forward pass. Conversely, during inference, these states are generated autoregressively on the fly using the standard PersonaPlex inference loop.

Hybrid system prompt alignment. PersonaPlex requires a hybrid system prompt (text and voice) to initialize its auto-regressive generation. To match this distribution of the pre-trained PersonaPlex at training time, we explicitly prepend a constructed system prompt to each Seamless clip before extracting the hidden states. Please refer to the supplement for more details.

4.2 Full-duplex motion tower

The motion tower is a causal decoder-only Transformer with L_m=32 blocks (one per PersonaPlex layer), dimension d_m=1024, h_m=16 self-attention heads of head dimension d_m/h_m=64, RoPE [32] self-attention, and SwiGLU [31] feed-forward layers. A single shared embedding $\mathbf{E} \in \mathbb{R}^{V \times d_m}$ maps

token ids to features, where the vocabulary $V = V_{\text{mot}} + 2$ includes two special speaker tags [A], [B] (see Fig. 2).

Self-attention with dyadic interleaving. We flatten the two motion streams into a single token sequence that alternates speaker tags and their K RVQ codes at every frame:

$$\mathbf{M} = \big[\ \underbrace{[\texttt{A}],\ \mathbf{m}_1^A,\ [\texttt{B}],\ \mathbf{m}_1^B}_{\text{frame 0}},\ \underbrace{[\texttt{A}],\ \mathbf{m}_2^A,\ [\texttt{B}],\ \mathbf{m}_2^B}_{\text{frame 1}},\ \dots\ \big]. \tag{1}$$

This arrangement yields a per-frame "step" length of $L_{\text{step}} = 2(K+1) = 46$ tokens. Consequently, applying causal self-attention over the unified sequence $\mathbf{M}$ simultaneously models three distinct dependencies: (i) the intra-frame coherence between a single speaker's K RVQ codes, (ii) the within-frame cross-speaker reactions ($A \to B$), and (iii) the long-range temporal dynamics across consecutive frames. Ours is the first architecture in which both sides of a dyadic conversation share a single autoregressive motion prior. Adding partner motion information dramatically improves the dyadic motion quality compared to our model without partner as measured in P-FD and Δ-User experiment in Tab 2.

Cross-attention with time-aligned speech-motion RoPE. Within the motion tower, each transformer block ℓ is paired one-to-one with a corresponding PersonaPlex block. These towers interact through a multi-head cross-attention sub-layer with h_c=12 heads of per-head dimension d_c=64, where queries are derived from the interleaved motion stream, and keys/values are computed from the frozen speech hidden states $\mathcal{H}_\ell$. Specifically, at each block ℓ, the motion-tower hidden state $\mathbf{h}_t \in \mathbb{R}^{d_m}$ at flattened token position t in $\mathbf{M}$ is projected by a trainable matrix $\mathbf{W}_q^\ell \in \mathbb{R}^{d_m \times h_c d_c}$ to produce the query $\mathbf{q}_t$. Similarly, trainable matrices $\mathbf{W}_k^\ell, \mathbf{W}_v^\ell \in \mathbb{R}^{d_s \times h_c d_c}$ project $\mathcal{H}_\ell$ into keys $\mathbf{K}_\ell$ and values $\mathbf{V}_\ell$. This learned-projection architecture provides the motion tower with the capacity to remap speech features from the pre-trained PersonaPlex—which were natively optimized for audio synthesis—into representations tailored for gesture generation.

Crucially, to temporally align the motion and speech modalities, we assign every motion token a query position $q_{\text{pos}}(t)$ corresponding to its actual frame index. Specifically:

$$q_{\text{pos}}(t) \ = \ \big\lfloor t \,/\, L_{\text{step}} \big\rfloor. \tag{2}$$

Consequently, all L_{step}=46 tokens comprising a single motion frame share the exact same query position (e.g., all tokens in the 0-th frame of $\mathbf{M}$ are assigned $q_{\text{pos}} = 0$). We then apply Rotary Positional Embeddings (RoPE [32]) to the queries and keys using these positions:

$$\begin{aligned} \tilde{\mathbf{q}}_t &= \mathrm{RoPE}\big(\mathbf{q}_t,\ q_{\text{pos}}(t)\big), \\ \tilde{\mathbf{k}}_s &= \mathrm{RoPE}\big(\mathbf{K}_\ell[s],\ s\big). \end{aligned} \tag{3}$$

where s denotes the temporal index of the speech tokens. Since the motion sampling rate matches the speech sampling rate ($f_m = f_s$), the indices $q_{\text{pos}}(t)$ and s operate on a *unified* temporal axis. For each block ℓ, we define $\tilde{\mathbf{K}}_\ell = [\tilde{\mathbf{k}}_1, \dots, \tilde{\mathbf{k}}_T]$ as the full sequence of rotated keys obtained by applying RoPE to each projected speech state. The cross-attention output for the ℓ-th layer is then calculated as $\mathrm{XAttn}^\ell(\mathbf{h}_t) = \mathrm{Attention}\big(\tilde{\mathbf{q}}_t,\ \tilde{\mathbf{K}}_\ell,\ \mathbf{V}_\ell\big)$. Since the RoPE computes attention based strictly on the relative offset $q_{\text{pos}}(t) - s$, the ideal solution simply reduces to a diagonal alignment, where each motion frame attends directly to its concurrent speech frame. This provides an inductive bias toward time-aligned attention, which the network learns to exploit during training. We refer to this mechanism as *time-aligned speech-motion RoPE*. Without RoPE, the cross-attention has no explicit positional signal and would have to recover the alignment implicitly through the motion query and speech key alone, which results in suboptimal speech-motion alignment. The importance of both cross attention and this speech-motion alignment via RoPE is empirically validated by the clear degradation in the BeatAlign score when they are removed (see Tab. 2; further analysis is provided in the supplement).

Speech context window and causality. To ensure strict causality for real-time streaming, the cross-attention mechanism dictates that a motion token t can only attend to speech frames at or preceding its concurrent motion frame. This constraint is enforced via a causal mask $M_{t,s}$:

$$M_{t,s} \ = \ \begin{cases} 1, & \text{if } s \le q_{\text{pos}}(t) \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

Theoretically, cross-attention can access an indefinite speech history without added motion self-attention overhead. For simplicity, we align the speech context with the motion tower's 4096-token window, bounding the receptive field to 89 frames (≈ 7.1 s at 12.5 fps).

4.3 Training Objective

We optimize the motion tower parameters θ via teacher-forced next-token prediction over the interleaved sequence $\mathbf{x}$, conditioned on the precomputed speech states $\{\mathcal{H}_\ell\}$. To prevent the model from assigning probability mass to structurally invalid tokens across the four disjoint RVQ codebooks, we apply a band-mask (setting out-of-band logits to $-\infty$) prior to the softmax. The masked cross-entropy loss is:

$$\mathcal{L}_{\text{CE}}(\theta) = -\sum_{t \in \mathcal{S}} \log p_\theta(x_{t+1} | x_{\le t}; \{\mathcal{H}_\ell\}),$$

where $\mathcal{S}$ denotes the 18 supervised body code positions per frame using the official SMPL-H body data (body + hands) that comes with the Seamless [1] dataset. In this paper we focus on body motion generation; throughout, "Ours" refers to this *body-only* base model. A separate variant of our model that also predicts face codes is used only for the qualitative demos in Fig. 1 and the supplementary video. Additionally, following MIBURI [22], we employ a linear voice-activation head to predict the binary speaking/listening state v_t at each valid code position $t \in \mathcal{V}$. This yields an auxiliary binary cross-entropy loss $\mathcal{L}_{\text{VA}}$. Our final training objective is simply $\mathcal{L} = \mathcal{L}_{\text{CE}} + \beta \mathcal{L}_{\text{VA}}$, where we empirically set $\beta = 0.01$.

4.4 Streaming inference

During inference, the model operates as a causal streaming sampler. The PersonaPlex speech tower synthesizes agent B's speech while also generating the per-layer hidden states $\mathcal{H}_\ell$ that encode the dyadic conversational context. Conditioned on $\mathcal{H}_\ell$, the motion tower can either generate both speakers' motion jointly or only the agent's motion, depending on the application. In the *both-speaker mode* (e.g., synthetic dyadic interaction data generation), the motion tower autoregressively samples motion tokens at both [A] and [B] slots, producing complete dyadic motion for both speakers. In the *agent-only mode* (e.g., human-agent / robot interaction), given an observed partner-motion prefix $\mathbf{m}^A_{1:f}$ and the PersonaPlex hidden states $\mathcal{H}_\ell$ for both speakers up to frame f, we fill in the observed partner tokens into the [A] slots and sample from the learned distribution p_θ exclusively at the agent's [B] slots:

$$\hat{c}^{(k)}_f \sim \text{topk}\big(\text{softmax}(\text{logits}(\mathbf{x}_{<t})/\tau),\ K_{\text{top}}\big), \tag{5}$$

where $\mathbf{x}_{<t}$ denotes the partial interleaved sequence up to flat position t (Eq. (1)). To rigorously maintain the dyadic structure, we deterministically insert the appropriate speaker tags ([A] or [B]) and copy the ground-truth partner motion $\mathbf{m}^A_f$ at their designated positions. The autoregressive sampling is restricted entirely to the [B] code positions, applying a temperature of $\tau = 1.0$ and $K_{\text{top}} = 200$ throughout.

Crucially, because the speech hidden states $\mathcal{H}_\ell$ are continuously produced by a frozen streaming speech tower, and the part-aware RVQ decoders (Sec. 3.2) are inherently causal, the entire generation pipeline—from partner audio input, through PersonaPlex and the motion tower, down to the final SMPL-X reconstruction—maintains strict causality. This architectural guarantee allows the system to be executed chunk-wise, enabling genuine real-time streaming interaction.

5 Experiments

Datasets. We utilize the Seamless Interaction dataset [1], a large-scale corpus featuring approximately 4000 h of dyadic conversations. After applying filters to remove invalid and corrupted data, 57 947 pairs (3435 h) of dyadic motion are available for training. For evaluation, we further restrict to a 330-pair test subset ($\sim$18 h), retaining only pairs where both speakers pass additional audio quality check (e.g., broken or missing recordings). We used the first 20 seconds of the test set for the evaluations. Both audio and motion are tokenized at a synchronized rate of 12.5 Hz using Mimi [4] and our causal RVQ-VAE, respectively. For details on data processing pipeline, please refer to the supplement.

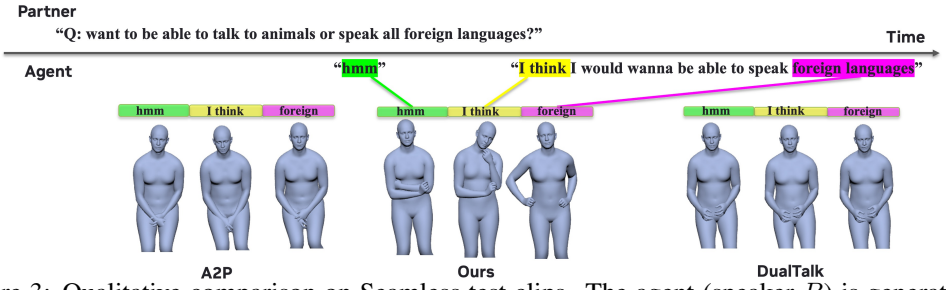

Figure 3: Qualitative comparison on Seamless test clips. The agent (speaker B) is generated by a different method, conditioned on the same ground-truth user (speaker A) motion and speech. Compared with Audio2Photoreal [24] and DualTalk [29] (frozen due to mode collapse), our model produces more diverse and natural dyadic behaviors.

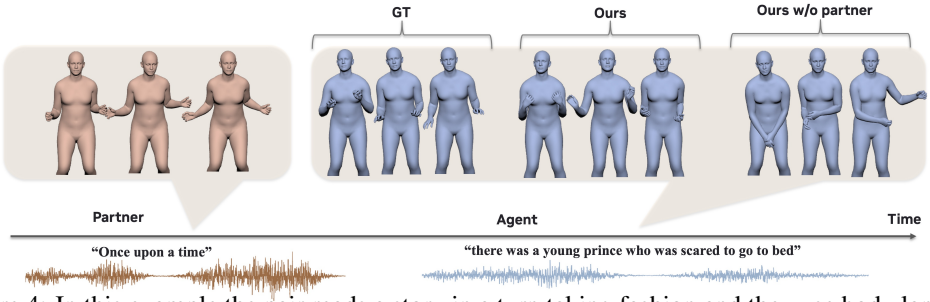

Figure 4: In this example the pair reads a story in a turn taking fashion and they use body languages to communicate turn taking. Our dyadic model produces more coherent gestures with the partner and the ground truth agent. Our model without perceiving the partner motion (Ours w/o partner) fails to produce coherent gestures as it cannot see the partner gestures.

Baselines. We compare our method with Audio2Photoreal [24] (A2P) and DualTalk [29], state-of-the-art dyadic conversational models using official source code. A2P is representative for a diffusion-based method and DualTalk for a transformer-based method. We adapt both baselines and retrain all methods, including our body-only base model, on the same Seamless SMPL body dataset (see the supplement for more details).

Metrics. We report two groups of metrics on the Seamless test set, summarized in Tab. 2: (i) *Monadic & Alignment Metrics*: FGD [37] and Diversity [24] evaluate individual motion quality, while BeatAlign [12, 15] measures speech-motion synchronization. (ii) *Dyadic Interaction Metrics*: Paired Fréchet Distance (P-FD) [23, 20] evaluates the realism of the joint user–agent motion distribution. User-Conditioning Gain (Δ-User) quantifies the model's reliance on partner motion, defined as the relative P-FD improvement when conditioning on matched versus shuffled user motions: $\Delta_{\text{User}} = (\text{P-FD}_{\text{shuffled}} - \text{P-FD}_{\text{matched}})/\text{P-FD}_{\text{matched}}$ expressed as a percentage. By definition, baselines lacking a user-motion pathway inherently score 0%. Since A2P and DualTalk uses ground-truth audio of partner and agents as input, for a fair comparison, we evaluate our model under a *teacher-forced* configuration: ground-truth PersonaPlex hidden states and user motion tokens are provided, and only the agent's motion tokens are sampled. All evaluations are conducted at 25 fps. Ground-truth body joints are derived from the Seamless SMPL-H parameters via SMPL-X forward kinematics. For evaluation, following Seamless [1], we set the root translation to zero for all methods. All metrics evaluate body motion and interaction quality using the 22 SMPL-X body joints (excluding finger and face joints).

5.1 Qualitative Results

Fig. 3 shows qualitative comparisons to the baselines. Our model produces gestures that are diverse and appropriate to the conversation while baselines generate less diverse results. Fig. 4 shows qualitative comparisons to our ablation model without partner perception (*w/o Partner*). Our model without partner perception fails to mimic partner gestures.

Table 2: **Quantitative results on Seamless.** Two column groups: Monadic metrics (left) and dyadic metrics (right). Columns marked "↓" are lower-is-better (FGD, P-FD); all other columns report values for which closer to GT is better. Δ-User cells with positive percentages indicate methods whose generated motion changes when user motion is shuffled at inference. † indicates methods do not use partner motion information as input and thus 0% by design. **Bold** and underline indicate best and second-best per column among baselines and Ours variants. DualTalk reports substantially worse results due to training collapse that results in frozen motion, and their BeatAlign is undefined due to no detectable motion.

Method	Monadic			Dyadic	
	FGD ↓ ($\times 10^{-3}$)	Diversity →GT	BeatAlign →GT	P-FD ↓ ($\times 10^{-3}$)	Δ-User ↑
GT	—	0.633	0.049	—	—
GT (Random)	13	0.683	0.050	33	0%†
Baselines					
Audio2Photoreal [24]	57	0.395	**0.051**	72	0%†
DualTalk [29]	161	0.305	—	163	+0.3%
Ablations (Ours)					
w/o Self-Attn	41	0.416	0.132	45	0%†
w/o Partner	39	0.725	0.064	41	0%†
w/o Cross-Attn	41	0.708	0.080	44	+15%
w/o Cross-RoPE	8.4	0.582	0.064	10	**+31%**
Ours (body-only)	**5.6**	**0.611**	0.059	**7.3**	**+31%**

5.2 Quantitative Results

Tab. 2 shows quantitative results against baselines. We additionally report ground truth (GT) and randomly sampled ground truth data (GT (Random)) as additional references. The elevated P-FD in GT (Random) confirms that P-FD works as expected to measure the joint motion distributions of two people. Our method significantly outperforms all baselines in both monadic and dyadic motion quality (FGD, P-FD). Our method has the most similar diversity to GT and scores second best BeatAlign.

Ablation study. Tab. 2 also shows comparisons to our ablation models. Removing self attention (*w/o self-attention*) leads to significantly worse results across the board. In the dyadic setting, removing partner perception (*w/o Partner*) degrades FGD $\sim 7\times$ ($5.6 \rightarrow 39$) and P-FD $\sim 5.6\times$ ($7.3 \rightarrow 41$) above the full model. Removing cross attention (*w/o cross-attention*) and RoPE (*w/o cross-attn RoPE*) leads to degraded BeatAlign scores due to entirely missing the speech context or imprecise alignment between speech-motion tokens. Δ-User shows how much P-FD a model gains when partner motion is available. As shown in +31% gain in P-FD in our method, in the dyadic setting, having the partner motion as input is crucial.

User Study. To evaluate perceived motion naturalness, we conducted a user study comparing our method against DualTalk [29] and Audio2Photoreal [24] on the Seamless test set [1]. Thirty-two participants used an interactive UI to compare paired mesh-rendered videos with the same ground-truth conversation audio; only the agent body motion differs between methods. As detailed in Tab. 3, our method is overwhelmingly preferred over DualTalk (97.5%) and Audio2Photoreal (66.3%), validating its superior generation quality. Notably, our generated motions achieve a 29.4% preference rate even when compared directly against the ground truth.

Table 3: **User study of naturalness of the generated agent motion on the Seamless [1] test set**. We report the percentage of the participants that prefer our method against the counterparts.

Comparison	Preference to Ours
vs. Ground Truth	29.4%
vs. Audio2Photoreal [24]	66.3%
vs. DualTalk [29]	97.5%

Runtime performance. Evaluated on a single RTX A6000 Ada GPU at 12.5 Hz, the audio tower and RVQ-VAE decoder run efficiently at 30ms and 0.8ms per frame, respectively. The autoregressive motion tower is the primary computational bottleneck, taking 173ms/frame with a full 4096-token

context. By reducing the motion context to 1024 tokens (1.8s) while preserving the full 7.1s speech context, the motion tower latency drops to 80ms/frame, successfully achieving real-time inference.

6 Conclusion

We introduced DyaPlex, a full-duplex speech-and-motion framework for streaming dyadic interaction. By coupling a frozen PersonaPlex speech tower with a trainable motion tower via our time-aligned speech-motion RoPE mechanism, we achieve precise cross-modal synchronization while maintaining strict causality. This approach elegantly reduces temporal alignment to a structural inductive bias that the motion tower learns to exploit during training. Trained on the 4,000-hour Seamless Interaction dataset, our model establishes new state-of-the-art performance in both individual motion realism and joint dyadic interaction. By ensuring end-to-end causal streaming, DyaPlex enables genuine, low-latency interactions, providing a scalable and efficient foundation for next-generation embodied conversational agents.

Acknowledgments and Disclosure of Funding

We thank David Luebke, Slim Essid, Nikhil Srihari, and Viet Anh Trinh for early discussions on the project.

References

[1] Vasu Agrawal, Akinniyi Akinyemi, Kathryn Alvero, Morteza Behrooz, Julia Buffalini, Fabio Maria Carlucci, Joy Chen, Junming Chen, Zhang Chen, Shiyang Cheng, Praveen Chowdary, Joe Chuang, Antony D'Avirro, Jon Daly, Ning Dong, Mark Duppenthaler, Cynthia Gao, Jeff Girard, Martin Gleize, Sahir Gomez, Hongyu Gong, Srivathsan Govindarajan, Brandon Han, Sen He, Denise Hernandez, Yordan Hristov, Rongjie Huang, Hirofumi Inaguma, Somya Jain, Raj Janardhan, Qingyao Jia, Christopher Klaiber, Dejan Kovachev, Moneish Kumar, Hang Li, Yilei Li, Pavel Litvin, Wei Liu, Guangyao Ma, Jing Ma, Martin Ma, Xutai Ma, Lucas Mantovani, Sagar Miglani, Sreyas Mohan, Louis-Philippe Morency, Evonne Ng, Kam-Woh Ng, Tu Anh Nguyen, Amia Oberai, Benjamin Peloquin, Juan Pino, Jovan Popovic, Omid Poursaeed, Fabian Prada, Alice Rakotoarison, Alexander Richard, Christophe Ropers, Safiyyah Saleem, Vasu Sharma, Alex Shcherbyna, Jia Shen, Jie Shen, Anastasis Stathopoulos, Anna Sun, Paden Tomasello, Tuan Tran, Arina Turkatenko, Bo Wan, Chao Wang, Jeff Wang, Mary Williamson, Carleigh Wood, Tao Xiang, Yilin Yang, Zhiyuan Yao, Chen Zhang, Jiemin Zhang, Xinyue Zhang, Jason Zheng, Pavlo Zhyzheria, Jan Zikes, and Michael Zollhoefer. Seamless interaction: Dyadic audiovisual motion modeling and large-scale dataset. 2025. URL https://ai.meta.com/research/publications/seamless-interaction-dyadic-audiovisual-motion-modeling-and-large-scale-dataset/.

[2] Zhi Cen, Huaijin Pi, Sida Peng, Qing Shuai, Yujun Shen, Hujun Bao, Xiaowei Zhou, and Ruizhen Hu. Ready-to-react: Online reaction policy for two-character interaction generation. In *ICLR*, 2025.

[3] Riccardo Corvi, Davide Cozzolino, Ekta Prashnani, Shalini De Mello, Koki Nagano, and Luisa Verdoliva. Seeing what matters: Generalizable ai-generated video detection with forensic-oriented augmentation. In *NeurIPS*, 2025.

[4] Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*, 2024.

[5] Panagiotis P Filntisis, George Retsinas, Foivos Paraperas-Papantoniou, Athanasios Katsamanis, Anastasios Roussos, and Petros Maragos. Visual speech-aware perceptual 3d facial expression reconstruction from videos. *arXiv preprint arXiv:2207.11094*, 2022.

[6] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. Remos: 3d motion-conditioned reaction synthesis for two-person interactions. In *European Conference on Computer Vision (ECCV)*, 2024.

[7] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14783–14794, October 2023.

[8] Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Lingjie Liu, Hans-Peter Seidel, Gerard Pons-Moll, Mohamed Elgharib, and Christian Theobalt. Learning speech-driven 3d conversational gestures from video. In *Proceedings of the 21st ACM international conference on intelligent virtual agents*, pages 101–108, 2021.

[9] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36:20067–20079, 2023.

[10] Wentao Jiang, Jingya Wang, Haotao Lu, Kaiyang Ji, Baoxiong Jia, Siyuan Huang, and Ye Shi. Arflow: Human action-reaction flow matching with physical guidance. *ArXiv*, 2025.

[11] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, et al. Panoptic studio: A massively multiview system for social interaction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16, 2017.

[12] Ruilong Li, Sha Yang, David A. Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *ICCV*, 2021.

[13] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017.

[14] Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion generation under complex interactions. *International Journal of Computer Vision*, 132(9): 3463–3483, 2024.

[15] Haiyang Liu, Zihao Zhu, Naoya Iwamoto, Yichen Peng, Zhengqing Li, You Zhou, Elif Bozkurt, and Bo Zheng. Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In *ECCV*, 2022.

[16] Haiyang Liu, Zihao Zhu, Giorgio Becherini, Yichen Peng, Mingyang Su, You Zhou, Xuefei Zhe, Naoya Iwamoto, Bo Zheng, and Michael J Black. Emage: Towards unified holistic co-speech gesture generation via expressive masked audio gesture modeling. In *CVPR*, 2024.

[17] Pinxin Liu, Luchuan Song, Junhua Huang, Haiyang Liu, and Chenliang Xu. Gesturelsm: Latent shortcut based co-speech gesture generation with spatial-temporal modeling. *arXiv preprint arXiv:2501.18898*, 2025.

[18] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[19] Thomas Lucas, Fabien Baradel, Philippe Weinzaepfel, and Grégory Rogez. Posegpt: Quantization-based 3d human motion generation and forecasting. In *European Conference on Computer Vision*, pages 417–435. Springer, 2022.

[20] Vongani H Maluleke, Lea MÃ¼ller, Jathushan Rajasegaran, Georgios Pavlakos, Shiry Ginosar, Angjoo Kanazawa, and Jitendra Malik. Synergy and synchrony in couple dances. *arXiv preprint arXiv:2409.04440*, 2024.

[21] M Hamza Mughal, Rishabh Dabral, Merel CJ Scholman, Vera Demberg, and Christian Theobalt. Retrieving semantics from the deep: an rag solution for gesture synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16578–16588, 2025.

[22] M. Hamza Mughal, Rishabh Dabral, Vera Demberg, and Christian Theobalt. Miburi: Towards expressive interactive gesture synthesis. In *CVPR*, 2026.

[23] Evonne Ng, Hanbyul Joo, Liwen Hu, Hao Li, Trevor Darrell, Angjoo Kanazawa, and Shiry Ginosar. Learning to listen: Modeling non-deterministic dyadic facial motion. *CVPR*, 2022.

[24] Evonne Ng, Javier Romero, Timur Bagautdinov, Shaojie Bai, Trevor Darrell, Angjoo Kanazawa, and Alexander Richard. From audio to photoreal embodiment: Synthesizing humans in conversations. In *CVPR*, 2024.

[25] Evonne Ng, Siwei Zhang, Zhang Chen, Michael Zollhoefer, and Alexander Richard. Sarah: Spatially aware real-time agentic humans, 2026. URL https://arxiv.org/abs/2602.18432.

[26] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019.

[27] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *CVPR*, 2024.

[28] Yichen Peng, Jyun-Ting Song, Siyeol Jung, Ruofan Liu, Haiyang Liu, Xuangeng Chu, Ruicong Liu, Erwin Wu, Hideki Koike, and Kris Kitani. Dyadit: A multi-modal diffusion transformer for socially favorable dyadic gesture generation. In *CVPR*, 2026.

[29] Ziqiao Peng, Yanbo Fan, Haoyu Wu, Xuan Wang, Hongyan Liu, Jun He, and Zhaoxin Fan. Dualtalk: Dual-speaker interaction for 3d talking head conversations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21055–21064, 2025.

[30] Rajarshi Roy, Jonathan Raiman, Sang gil Lee, Teodor-Dumitru Ene, Robert Kirby, Sungwon Kim, Jaehyeon Kim, and Bryan Catanzaro. Personaplex: Voice and role control for full duplex conversational speech models, 2026. URL https://arxiv.org/abs/2602.06053.

[31] Noam Shazeer. Glu variants improve transformer, 2020.

[32] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomput.*, 2024.

[33] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *European Conference on Computer Vision*, pages 358–374. Springer, 2022.

[34] Zhenzhi Wang, Jingbo Wang, Yixuan Li, Dahua Lin, and Bo Dai. Intercontrol: Zero-shot human interaction generation by controlling every joint. *Advances in Neural Information Processing Systems*, 37:105397–105424, 2024.

[35] Liang Xu, Yizhou Zhou, Yichao Yan, Xin Jin, Wenhan Zhu, Fengyun Rao, Xiaokang Yang, and Wenjun Zeng. Regennet: Towards human action-reaction synthesis. In *CVPR*, 2024.

[36] Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. Generating holistic 3D human motion from speech. In *CVPR*, 2023.

[37] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM TOG*, page 1–16, 2020.

[38] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[39] Juze Zhang, Changan Chen, Xin Chen, Heng Yu, Tiange Xiang, Ali Sartaz Khan, Shrinidhi Kowshika Lakshmikanth, and Ehsan Adeli. Vibes: A conversational agent with behaviorally-intelligent 3d virtual body. In *CVPR*, 2026.

[40] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, 2019.

[41] Lingting Zhu, Xian Liu, Xuanyu Liu, Rui Qian, Ziwei Liu, and Lequan Yu. Taming diffusion models for audio-driven co-speech gesture generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10544–10553, 2023.

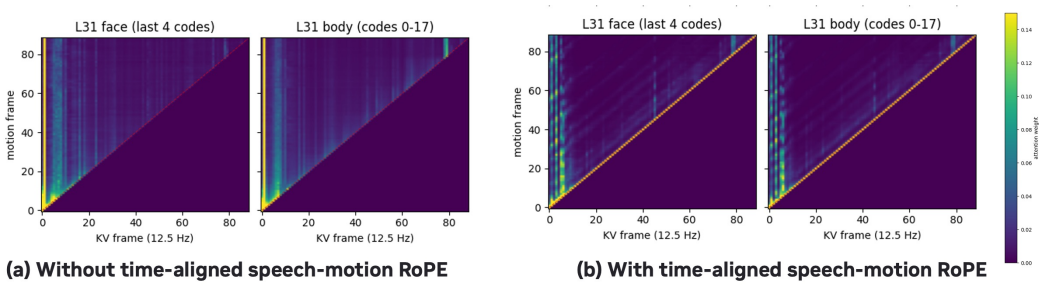

Figure 5: Comparisons without (a) and with (b) time-aligned speech-motion RoPE.

A Discussion

A.1 Limitations and future work

Our model currently decodes one body token at a time for 22 codes, which may not be most optimal for performance. For the actual deployment in robot, it would be worthwhile to explore chunk by chunk decoding (e.g., 6 upper body token in one go) or the decomposed depth transformer-based design similar to Moshi [4] to speed up the inference. It would be also fruitful future work to explore other generative models backbone such as diffusion models. Our model can handle interactions with only 2 users. Extending our model to polyadic interaction would be interesting future work.

A.2 Potential negative societal impacts

DyaPlex outputs generic SMPL-X body pose parameters rather than photorealistic pixels, so the model itself does not directly produce identity-bearing video. However, generated motion could be paired with downstream identity-conditioned video diffusion models to produce realistic-looking videos of fabricated dyadic interactions, creating potential risks for accessible deepfakes. We highlight a few mitigations that follow naturally from the design of DyaPlex. First, our model is *identity-agnostic*: it generates motion on the generic SMPL-X skeleton without conditioning on any subject identifier, so impersonation of a specific person requires a separate identity-conditioned avatar/voice pipeline whose own safeguards apply. Additionally, there exist tools to detect videos generated by AI models based on pixel-footprints intrinsic to video generators (e.g., [3]). Second, the speech tower (PersonaPlex) is frozen and used in its public release configuration; we add no fine-tuning that targets specific real-world voices. On the positive side, the same capability supports beneficial applications: scalable synthetic dyadic interaction data for training social robots and embodied agents, and virtual agents and robots for full-duplex human AI interactions as discussed in the introduction.

B Effect of time-aligned speech-motion RoPE

Fig. 5 ablates the importance of having time-aligned speech-motion RoPE in the cross attention weight. The diagonal dotted red line shows an expected alignment. The figure shows that without RoPE the motion cannot cleanly attend to the speech features at the same time frame. With RoPE, the cross attention map lights up diagonally, demonstrating the motion tokens are correctly attending to the speech feature of that frame.

C Details of Seamless dataset processing

C.1 Dataset filtering

Seamless Interaction [1] ships per-clip SMPL-H body parameters recovered by per-frame HMR2 [7] regression and per-frame hand parameters by HaMeR [27]. Both are off-the-shelf monocular regressors; their failure modes leak into training unless filtered explicitly. We apply two filters: a clip-level aspect-ratio filter and a frame-level validity mask.

Aspect-ratio and integrity filter (clip level). The Seamless videos cover a heterogeneous mix of aspect ratios. Roughly 94.5% are typical $9{:}16$ portrait clips, with smaller fractions of square and near-square crops; the remaining $\sim 1.5\%$ are landscape (3840×2160, 1920×1080), rotated (640×480), or zero-resolution unreadable / corrupt files. HMR2 produces unusable body fits on landscape or rotated frames and fails outright on unreadable ones. We filter at the pair level: a dyadic pair is dropped if either speaker's video is flagged as landscape, rotated, or unreadable, removing approximately 1.5% of the pairs in the dataset.

Per-frame HMR2 validity (frame level). Even on aspect-correct videos, HMR2 fails to detect a body in $\sim 2.4\%$ of frames. The dataset's SMPL-H NPZs carry a per-frame `smplh:is_valid` flag marking these. We drop frames where `is_valid == 0` from all downstream processing (statistics computation, RVQ-VAE training input, motion-tower training input, evaluation). Pairs with fewer than 8 valid frames after masking are dropped entirely.

C.2 Adding facial expressions

Seamless Interaction's body annotations are SMPL-H (body + hands only) and do not include facial expression parameters. For preliminary face experiments we augment the dataset with per-frame FLAME [13] face parameters extracted by SPECTRE [5] run on the source videos. SPECTRE is a video-conditioned monocular face reconstruction model that produces per-frame FLAME jaw rotation, expression coefficients, and shape coefficients.

Face representation. For each Seamless video, SPECTRE outputs per-frame 6D head rotation, 6D jaw rotation, and 50D FLAME expression PCs. We retain the 6D jaw rotation and the 50D expression to form a 56-dimensional face input, discarding the 6D head rotation (redundant with the SMPL-H body's joint 0, the global root).

SPECTRE quality filter. SPECTRE produces all-zero face vectors on frames where face detection or fitting fails. We treat any pair with $\geq 20\%$ zero-face frames in either speaker as unusable for face training and exclude it from the face-codec training subset. Remaining isolated zero frames within accepted pairs are kept as-is.

Scope. Face supervision in the Seamless dataset is preliminary. The extracted-FLAME corpus is used only for the independent RVQ-VAE encoder and decoder corresponding to the face (Sec. D.3) (RVQ-VAE encoders and decoders are independent per body part) and a variant of motion tower which jointly predicts body, hands and a face and only used in the qualitative face demos in Fig. 1 and the supplementary video.

D Implementation details

D.1 Details of PersonaPlex hybrid prompt in Seamless

We feed Seamless data into PersonaPlex's training-time `forward_train()` using the same hybrid system prompt system as PersonaPlex [30] and agent inner-monologue text format as Moshi [4]. The text system prompt itself is unchanged from the PersonaPlex default for unstructured conversations (`"<system> You enjoy having a good conversation. <system>"`). The voice prompt and inner-monologue text is pre-processed for the Seamless dataset as follows:

Voice system prompt. PersonaPlex's hybrid prompt requires a ~ 10 s pre-tokenized voice clip per agent (8 Mimi codebooks) to prime the model on the agent's voice identity. For each Seamless agent we use a clean ~ 10 s voice sample of that speaker, Mimi-tokenized in advance. Roughly 1.4% of pairs are dropped because no voice clip is available for one of the two participants.

Agent inner-monologue text stream. PersonaPlex predicts a text stream along with voice stream at matching 12.5Hz token rate. The agent side of the word-aligned seamless conversation transcripts are preprocessed to the dense token format as in PersonPlex and Moshi. In this format, sub-word tokens for a word are temporally aligned to the beginning of the word's utterance in the audio stream, and remaining frames are filled with PAD and EPAD tokens.

D.2 Additional Details of Motion Tower

This subsection covers the training-time hyperparameters of the motion tower; the architecture and loss are described in Sec. 4.

Training details. We use AdamW [18] optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay 0.1, gradient clipping at ℓ_2-norm 1.0, and learning rate 3×10^{-4}, effective batch size 512, and train our model on 64 NVIDIA H100 GPUs for 30K iterations.

Inference. We employ streaming autoregressive sampling with temperature $\tau = 1.0$ and no top-k truncation.

D.3 Details of causal RVQ-VAE

We adapt the part-aware RVQ-VAE of GestureLSM [17] into a streaming codec for Seamless Interaction [1]. This appendix covers per-part input dimensions, the streaming architecture changes, and how the four decoded outputs assemble into a complete SMPL-X pose at inference.

Per-part decomposition. Seamless body annotations follow the SMPL-H convention; we extract SMPL-X [26]body+face parameters by joint-set re-mapping and combine the result with SPECTRE [5] -extracted FLAME [13]face parameters. The four part-aware RVQ-VAEs operate on disjoint slices of this combined representation:

Part	Input dim	Joints / fields	K quantizers	Codebook band
Upper	78	13 joints $\{3, 6, 9, 12{-}21\}\times$ 6D rot	$K_b = 6$	$[0, 1024)$
Hands	180	30 joints $\{25{-}54\}\times$ 6D rot	$K_b = 6$	$[1024, 2048)$
Lower + trans	57	9 joints $\{0, 1, 2, 4, 5, 7, 8, 10, 11\}\times$ 6D rot + 3D translation velocity	$K_b = 6$	$[2048, 3072)$
Face	56	6D jaw rot (joint 22) + 50D FLAME expression	$K_f = 4$	$[3072, 4096)$

Each part is independent at the codec level (separate encoder, decoder, codebook); the four streams are unified only at the motion-tower input where their token IDs are interleaved into the dyadic stream described in Sec. 4.2. The shared V_{mot}=4096 vocabulary is the union of the four disjoint 1024-entry bands.

Architecture modifications. We make only two changes to the public GestureLSM RVQ-VAE; every other encoder, decoder, and quantizer setting is inherited unchanged. **(i)** We make the encoder *causal* by padding each dilated convolution on the left only, so the code emitted at frame f depends on input frames up to f and never on future frames—giving zero look-ahead at deployment. **(ii)** We halve the temporal downsampling, from the original $4\times$ to $2\times$, using a single strided downsampling stage instead of two. Applied to motion at the 25 fps rate used for tokenization (below), this $2\times$ factor yields a 12.5 Hz token stream that aligns one-to-one with the 12.5 Hz speech features.

Training data. The four part-codecs are trained on Seamless Interaction body motion that has passed through the preprocessing of Sec. C.1: the clip-level aspect-ratio filter and the per-frame validity mask, translation rescaling and Savitzky–Golay smoothing, a fixed 180° rotation about the x-axis that maps the recovered poses into the SMPL-X world frame, and conversion to a 6D-rotation pose representation. The face codec is trained on the SPECTRE-extracted FLAME parameters; pairs in which either speaker has $\geq 20\%$ frames with failed face detection are excluded from face-codec training only.

Training recipe. Each part-codec is trained independently on a single NVIDIA H100 GPU, using the same optimizer and reconstruction objective as the public GestureLSM model, and is selected by held-out reconstruction error.

Tokenization. Body and face motion are resampled to 25 fps and encoded; with the $2\times$ downsampling this produces token streams at f_m=12.5 Hz. For motion-tower training the encoders are run once per training pair offline and the codes cached to disk as 16-bit integer bins of shape $(T, 22)$, so the training loop never re-invokes the codec. At inference the encoders run online on the partner's observed motion to produce the partner token stream that is teacher-forced into Eq. (5), while the agent's own tokens are sampled from the motion tower and decoded by the four part-decoders.

Output assembly and SMPL-X forward kinematics. Each decoder reconstructs the SMPL-X pose parameters of its joint subset. The four reconstructions are concatenated in canonical SMPL-X joint order (global orientation + 54 body / hand / face joints + translation + FLAME expression) and passed through the neutral SMPL-X model under forward kinematics with translation set to zero, matching the zero-translation evaluation protocol used for all body baselines. The output is per-frame body+face joint positions and mesh vertices.

E Details of Evaluations

E.1 Details of Metrics

All metrics are computed by our own evaluation pipeline; for the distributional metrics we reuse the reference Fréchet-distance implementation from ViBES [39].

FGD Fréchet distance on the agent's raw $22 \times 3 = 66$-D SMPL-X joint positions per frame, pooled across the test set. We adopt the FGD term from Yoon et al. [37], but compute the Fréchet distance directly on raw joint positions (no learned feature extractor), following the raw-geometry protocol of recent dyadic-gesture work (SARAH [25], DyaDiT [28]) rather than the learned autoencoder latent space of the original FGD.

BeatAlign Gaussian-Average Hit Rate between the agent's audio onsets and per-joint motion velocity-minima on the 13 upper-body joints, following the AIST++ [12] / BEAT [15] formulation with the audio-anchored alignment direction of ViBES [39] (for each audio onset, the distance to the nearest motion beat); audio onsets are extracted at Mimi's 24 kHz native sample rate.

Diversity Mean L2 distance between random pairs of agent-pose frames pooled across the test set, following the Audio2Photoreal [24] formulation (their `calculate_diversity`); closer to GT is better.

P-FD Paired Fréchet Distance [23, 20], computed on 132-D per-frame vectors formed by concatenating both speakers' 22×3 joint positions, pooled across the test set. We report the Fréchet distance between $(\mathrm{GT}_A, \mathrm{GT}_B)$ and $(\mathrm{GT}_A, \mathrm{Gen}_B)$ distributions.

Δ-User Relative gain in P-FD when each pair's user-motion track is replaced by a shuffled (mismatched) one; see main text for the formula.

E.2 DualTalk Implementation Details

We adapt DualTalk [29] from its original face-blendshape regression setting to SMPL-X body-pose regression on the Seamless Interaction dataset. We keep DualTalk's four-module architecture—(a) Dual-Speaker Joint Encoder, (b) Cross-Modal Temporal Enhancer, (c) Dual-Speaker Interaction Module, and (d) Expressive Synthesis Module—at its original depth (2 LSTM layers, 3 transformer encoder layers, 1 transformer decoder layer), together with the Wav2Vec 2.0 large audio backbone (`facebook/wav2vec2-large-960h-lv60-self`).[3] We make the following changes.

Output representation. We replace DualTalk's 56-D facial blendshape output (50 expression + 3 jaw + 3 neck) with a 69-D SMPL-X body parameterization (63-D body pose + 3-D global orientation + 3-D translation, all in axis-angle). The input motion encoder and the output projection head are resized accordingly.

[3] DualTalk public release: https://github.com/ziqiaopeng/DualTalk

Training. Each speaker's audio is encoded by a Wav2Vec 2.0 backbone, initialized from public pre-trained weights, with the convolutional feature extractor frozen and the remaining Wav2Vec 2.0 layers fine-tuned; all other components—the audio projection, the motion encoder, and modules (b)–(d)—are randomly initialized and trained. We use DualTalk's original objective, a sum of per-component mean-squared errors plus a frame-to-frame velocity term.[3] We optimize with Adam at a learning rate of 1×10^{-4} on 8 GPUs. The Seamless Interaction dataset (4,000 h) is roughly $80\times$ larger than the 50-hour dataset used in the original DualTalk work.[3]

Discussion. Trained on Seamless body data, DualTalk collapses to a near-constant body pose (per-frame body-velocity std $<10^{-5}$, vs. $\sim10^{-2}$ for our model), with the velocity term decreasing throughout training as the output converges to a static pose. We attribute this to two factors. (1) Under a mean-squared-error objective, the sparse, low-amplitude gestures in conversational body motion make the constant mean-pose solution a strong local optimum. (2) DualTalk's architecture and loss were designed for audio-to-face (lip-sync) generation, where the audio-to-motion mapping is near-deterministic and temporally dense; audio-to-body gesture is only weakly and non-deterministically coupled to speech, providing little reliable per-frame signal to learn—independent of how the model is initialized.

E.3 Audio2Photoreal Implementation Details

We adapt Audio2Photoreal [24], originally a full-body photorealistic-avatar synthesizer trained on Meta's proprietary high-fidelity dyadic capture, to our body-pose evaluation setting on the publicly released Seamless Interaction dataset. We keep the three-stage architecture—(a) a TemporalVertexCodec VQ-VAE that tokenizes body motion into a discrete codebook, (b) a Guide transformer that auto-regressively predicts keyframe tokens conditioned on audio, and (c) a FiLMTransformer diffusion model that denoises the dense per-frame motion conditioned on audio plus the Guide's keyframe tokens—and the public-release classifier-free guidance recipe[4]. We make the following changes.

Output representation. A2P natively predicts latent face-expression codes and kinematic-skeleton body joint angles; we replace this output with a 132-D SMPL-X body parameterization (22 joints in 6-D continuous rotation [40]). We subtract each clip's mean translation and run SMPL-X forward kinematics with zero global translation (pelvis at the origin), matching the zero-translation evaluation protocol. The model generates agent B's body from the conversational audio of both speakers; A2P is audio-driven and takes no motion input. Because our task is body-only, we discard the public release's face branch and retrain only the body pathway.

Architecture and training objective. We use A2P's three-stage architecture, model capacities, training losses, and optimizer settings exactly as released[4], including the frozen vq-wav2vec audio encoder. The only model-side change is for frame rate: the public code hard-codes the audio-feature length and conditioning positions for 30 fps motion, so at our 25 fps rate we recompute them once from the (deterministic, fixed-stride) wav2vec feature extractor for the target window length.

Data and training schedule. We train on the Seamless Interaction subset shared by our other baselines, using 24-second windows at 25 fps. The tokenizer (300 000 iterations) and Guide (30 000 steps) are each trained on a single GPU; for the diffusion model we depart from the single-GPU public recipe and train across 8 GPUs (per-GPU batch 4, an effective batch of 32) for 800 000 steps.

Inference. Inference follows the public recipe—chunked diffusion over consecutive 24-second windows with the released classifier-free guidance—after which we recover meshes by SMPL-X forward kinematics under the zero-translation convention above.

E.4 Details of User Study

To evaluate the perceived quality of the generated body motion, we designed a web-based interactive interface to playback and compare the result videos. An example is shown in Figure 6. Note that the two videos can be played independently so the participants can freely toggle between the two methods

[4] Audio2Photoreal public release: https://github.com/facebookresearch/audio2photoreal/

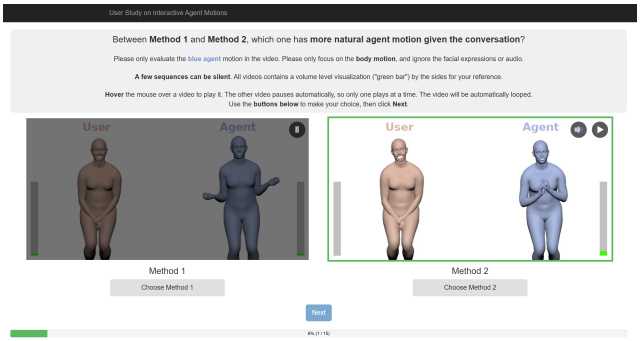

Figure 6: Interactive interface for the user study.

for evaluation and comparisons. In each result video, we render the animated mesh rendering of the partner (in peach color) and the generated agent (in blue color) along with the conversation audios. We randomly sample the comparison type (vs. the ground-truth, vs. Audio2Photoreal, vs. DualTalk) and randomly choose the display order on the web interface within the pair. Each pair of result videos shows the identical input partner and audio, with generated agent motion from different methods. The partner audio is encoded to the left audio channel and the generated agent audio is encoded to the right channel. Consequently, we require all participants to wear a headphone that supports stereo sound for this study. Since some input sequences contain near silent audio, we further visualize the live audio volume on the result videos, as a reference for the participants. During the evaluation, we ask the participants to focus on the body motion, and ignore the facial expressions or audio.